Application of Random Walks in Data Processing

Yujian Zhang [1,*], Hechen Zhang[2]

[1] Department of Mathematics, Liberal Arts and Sciences, University of Florida, Gainesville, 32603, United States

[2] University of Toronto, Toronto, Canada

*Corresponding Author Email: zhangyujian@ufl.edu

Abstract. A random walk is known as a process that a random walker makes consecutive steps in space at equal intervals of time and the length and direction of each step is determined independently. It is an example of Markov processes, meaning that future movements of the random walker are independent of the past. The applications of random walks are quite popular in the field of mathematics, probability and computer science. Random walk related models can be used in different areas such as prediction, recommendation algorithm to recent supervised learning and networks. It is noticeable that there are few reviews about randoms for the beginners and how random walks are used nowadays in distinctive areas. Hence, the aim of the article is to provide a brief review of classical random walks, including basic concepts and models of the algorithm and then some applications in the field of computer science for the beginners to understand the significance and future of random walks.

Keywords: Random walk, Non-backtracking random walk, Network Embedding.

1. Introduction

A random walk is a random process that describes a path that consists of consecutive random steps on some mathematical space. The most common example is of a random walk on the integer number line **Z**. The random walker starts from 0 and moves one unit in each time interval, meaning that +1 or -1 on the integer number line with equal probability. In reality, many phenomena contain the concept of random walks such as a molecule moving in liquid or gas, that is, Brownian motion. The concept of random walk was firstly introduced by Karl Pearson in 1905, in which can be categorized to an off-lattice random walk [1].

To understand an off-lattice random walk, a lattice walk is a good model to start with. In a lattice walk, the path of the random walker consists of connected horizontal and vertical line segments with each line passing two adjacent lattice points. Therefore, a lattice walk path is a sequence or set of points {P1, P2, …, Pn} with n>=0. Each Pi is a lattice point and Pi+1 is one unit away from last lattice point. In the meanwhile, the distance from Pn after n trials to the original point is calculated to study the correlation of these points according to a certain probability distribution.

In the two-dimensional scenario, there are four directions for a random walker to choose in the next step. Thinking about the random walker starting from the original point. The movement can be considered as x+, x-, y+, y- with equal probability, also know R2 in the Euclidean space. The random process can be obtained by generating a random number to determine the next step. Similarly, in the three-dimensional scenario, the directions of the random walker become six, plus z+ and z-, corresponding to the R3 Euclidean space. Therefore, the random process with a path consisting of continuous random steps on some mathematical space defines the random walk. Here we denote the random step as $\{\xi t: t = 0, 1, 2, \ldots\}$ where ξt is a random variable describing the position of a random walk after t steps [2]. The probability distribution for the position of the random walker only depends on the most recent position.

The applications of random walks could extend to many fields, including computer science, chemistry, physics, biology, sociology, economics, etc. In the field of computer science, random walk can be seen as a network topology and some algorithms recent years are based on random walks. For instance, random walks can be used in the link prediction and recommender system to calculate the

k-most-close nodes for the selected node in an effective way. Also, a high-level classification technique uses random walk theory to process input labeled data. Another example is "HRWN", hierarchical random walk network, which improves the accuracy of classification process from 88.91% to 93.61%, leading to an improvement of approximately 5% [3].

Besides, random walks are also important in the area of natural science. On the contrary to lattice random walk, the movements of particles in real world are not necessarily in the direction of a cardinal coordinate system. Instead, we use polar coordinate system to represent the movement of random walker in 2-dimension and spherical coordinate system in 3-dimensional. Topologically, Brownian motion is a real-life representation of random walk in 3-dimension since particles or molecules in a medium like liquid and gas moves randomly. The particles bounce in a medium and change direction every time they hit another particle. Random walk theory helps people calculate how long it will take to travel from a position to another. Consider a drop of ink spreading out in a glass of water is partially due to random walk. Apparently, real-life random processes of particles are much more complicated than the model of random walks. As a result, computer science and its algorithms can help people understand the random process in a more clear and reliable way.

2. Background

Here we denote the random step as $\{\xi_t: t = 0, 1, 2, \ldots\}$ where ξ_t is a random variable describing the position of a random walk after t steps. The probability distribution for the position of the random walker only depends on the most recent position. This stochastic process with memory no more than one step back is an example of Markov chain. The probability distribution of the position after t steps is:

$$Pt(i) = Pr(\xi_t = 1) \tag{1}$$

Where Pt(i) is the probability that the random walk visits the position i after t steps.

Another type of random walks in application is called non-backtracking random walk (NBW). It is simply a random walk that is conditioned not to go back to position or route that the random walker travelled, also known as random walk with no intersections or self-avoiding random walk (SAW). In mathematics, the self-avoiding random walk is a sequence of movements that does not visit the same point more than once [4].

Note that the non-backtracking random walk on a graph is not a Markov chain since at any point we need to remember the previous step to take the next step. In this random walk algorithm [7] on an edge-weighted directed graph $G = (V, E)$ with a weighting function on the set of edges $w: E \rightarrow \mathbb{R}^+ \cup \{0\}$, and the nonnegative weight of the directed edge ij between nodes i and j is represented by w_{ij}. Because of model limitations, only data locality improvements can be employed. Three data reordering metrics based on the p-sum functionals [8] with p=1, 2 and ∞ and the work bound functional [9] was tested [10]. Let π be a bijection $\pi: V \longrightarrow (1,2,\ldots,n)$. The following functional is minimized for the minimum p-sum problem over all possible permutations $\pi: \sigma_p(G, \pi) = \sum_{ij \in E} \left(w_{ij}|\pi(i) - \pi(j)|^p\right)^{1/p}$. The equation $bw(G) = \min_{\pi} \max_{ij \in E} w_{ij}|\pi(i) - \pi(j)|$ solves the minimum bandwidth problem when $p = \infty$, which finds a linear layout that minimizes the maximal stretched edge. The minimization functional of the workbound reduction problem is defined as $wb(G, \pi) = \sum_i \max_{\substack{j \\ \pi(j)<\pi(i)}} w_{ij}(\pi(i) - \pi(j))^2$.

3. Recent Researches

In this section, we revise some recent applications of the random walk algorithm.

Applying the random walk theory, the article proposed a supervised classification technique to process input label data. The random walk process over the adjacency matrix measures ease of access

and then estimate the class of a given unlabeled sample [5]. The advantage of the adjacency matrix using random walk process is that it carries both physical and structural information about the data. Hence, the user can put different data in the matrix of the network to guide the random walker. The new network-based data classification technique use labeled data as input with each instance being a network node. The labeled nodes are seen as pixels in random walk and assigned to an unlabeled node in the next step. Then, the training networks will classify the unlabeled instances one by one. At last, as the system's output, each of an unlabeled instance is given an estimated label by the most easily reached class through the network nodes, that is, an ease of access criterion. To measure the criterion, the limiting probabilities from random walk theory is used and computing the state transitions through nodes of the network. The other advantage using the limiting probabilities is that they take the whole network into account, considering the local and global relationships among data.

Literature [6] focused on the weak boundary and spatially fragmented classification issue and proposed an effective hierarchical random walk network, as known as "HRWN". Traditionally, to address spatial continuity, Markov random field (MRF) was integrated into the classification structure after the processing step. However, the model typically involves a great number of parameters and complex loss functions requiring specialized model training, making the training and testing quite complicated. As a result, the simple but flexible random walk method is used to improve the efficiency of the process. Statistically, the two branches CNN classifiers achieve an accuracy of 88.91% while the proposed HRWN classifier attains an accuracy of 93.61%, leading to an improvement of approximately 5%.

Literature [7] introduces a novel algorithm called "CARE" which can be used for different types of networks. However, the new CARE approach can preserve the community information of the network in the learned representation vector, which previous studies could not explicitly consider to define an optimization function. CARE preserves the characteristics of the network structure by generating custom paths for each node independently. Therefore, it spends less time learning the final representation of the nodes due to the parallel path generation. This paper empirically evaluates the algorithm for multi-label classification and link prediction problems in different realistic social networks. The experimental results show that the method has high embedding efficiency compared to other network embedding methods.

Literature [8] introduces a novel algorithm called "CARE", which can be used for different types of networks. CARE builds custom paths composed of local and global structures of network nodes as the basis for network embedding, and uses Skip-gram model to learn the representation vectors of nodes. In recent years, deep learning has been widely used to process natural language. Machine learning techniques are applied to extract valuable features from social networks. These techniques attempt to extract local structural information from each node and then use them to learn the final representation of the node. CARE can be extended without loss of information when new nodes are added to the network, and network embedding is an unsupervised representation learning task that attempts to extract information-rich, low-dimensional representations of network nodes. These unsupervised methods can handle the scalability of feature learning methods; however, the extracted features are too general to provide accurate information for a specific task. It learns the social relationships of network nodes in the low-dimensional space to maintain the microscopic and macroscopic network structure including various neighborhood orders, community members and their inherent properties. However, the new CARE approach can preserve the community information of the network in the learned representation vector, which previous studies could not explicitly consider to define an optimization function [9, 10]. CARE preserves the characteristics of the network structure by generating custom paths for each node independently. Therefore, it spends less time learning the final representation of the nodes due to the parallel path generation. This paper empirically evaluates the algorithm for multi-label classification and link prediction problems in different realistic social networks. The experimental results show that the method has high embedding efficiency compared to other network embedding methods.

4. Challenges

Although the random walk method is a relatively well-developed approach in the field of deep learning, most of the existing models are very general and do not provide accurate information for specific missions. Since a large number of models focus only on theoretical explanations, choosing a reasonable model for a specific situation is still an open problem. On the other hand, an accurate model that can be adapted to most situations should be considered.

With the rapid development and growth of the network size, another problem of random walk algorithms arises. After finding that fast random walk graph kernels could not fit in main memory, researchers divided the graph into several clusters to solve this problem; however, due to the huge size of the graph, these methods still have long delays. In addition, these approximation methods do not represent the complex structure of random walk models well or do not show the connectivity of nodes in the real world. For instance, Luxburg et al. [11] solved the problem finding the travel time for working with simple and valuable formula characterized by rapid calculation and high accuracy with two methods, measuring the flow parameters of electrical network and the spectral parameters. These two strategies apparently demonstrate that the approximation of the commute time is unrelated to the global properties of large sizes of graphs. As a result, existing estimation of commute time still leaves a legacy that cannot be perfectly applied in the reality with many complex situations that should be considered.

5. Conclusion

This paper presents the research on random walk theory and some of the applications of random walk in the areas of computer science and natural science. This paper first introduce the typical computing process of random walk in Euclidean space and the definition of Markov chain. Next, this paper shows how random walk method is applied in a variety of academic aspects, including calculating the selected k-most-close nodes efficiently, increasing the accuracy of classification process with hierarchical random walk network, predicting the path of movements of little particles, and calculating the time it takes a particle to move to another position.

Random walk is an efficient and convenient method that is able to use in a wide range of areas. With numerous studies, it has been proven that using random walk can better improve Although there are some limitations, researchers try to investigate more sophisticated theoretical and practical applications. Further research will be helpful to provide more general and accurate models to simulate the effective algorithms in realistic situations.

References

[1] K.Pearson, “Theproblemoftherandomwalk,”Nature,vol.72, no.1867, 1905, Art. no. 342.

[2] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan and X. Kong, "Random Walks: A Review of Algorithms and Applications," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 2, pp. 95-107, April 2020, doi: 10.1109/TETCI.2019.2952908.

[3] X. Zhao et al., "Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7355-7370, Oct. 2020, doi: 10.1109/TGRS.2020.2982064.

[4] Fitzner, R., van der Hofstad, R. Non-backtracking Random Walk. J Stat Phys 150, 264–284 (2013). https://doi.org/10.1007/s10955-012-0684-6

[5] T. H. Cupertino, M. Guimarães Carneiro, Q. Zheng, J. Zhang, and L. Zhao, “A scheme for high level data classification using random walk and network measures,” Expert Systems with Applications, vol. 92. Elsevier BV, pp. 289–303, Feb. 2018. doi: 10.1016/j.eswa.2017.09.014.

[6] X. Zhao et al., "Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7355-7370, Oct. 2020, doi: 10.1109/TGRS.2020.2982064.

[7] L. Lovasz. Random walks on graphs: A survey. Bolyai Soc. Math. Studies, 2: 1–46, 1993.

[8] Martin Juvan and Bojan Mohar. Optimal linear labelings and eigenvalues of graphs. Discrete Appl. Math., 36(2): 153–168, 1992.

[9] Gianna M. Del Corso and Francesco Romani. Heuristic spectral techniques for the reduction of bandwidth and work-bound of sparse matrices. Numerical Algorithms, 28(1–4): 117–136, December 2001.

[10] T. Davis. University of florida sparse matrix collection. NA Digest, 97(23), 1997.

[11] U. Von Luxburg, A. Radl, and M. Hein, "Hitting and commute times in large graphs are often misleading," 2010, arXiv: 1003.1266.